BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–7

Inferring Ancestral Genomes with Duplications

Liying Cui [1], *Jijun Tang* [2,*], *Bernard M. E. Moret* [3] *and Claude dePamphilis*[1]

[1] *Department of Biology, Institute of Molecular Evolutionary Genetics, and Huck Institute of Life Sciences, Pennsylvania State University, University Park, PA 16802, USA*
[2] *Department of Computer Science & Engineering, University of South Carolina, Columbia, SC 29208, USA*
[3] *Department of Computer Science, University of New Mexico, Albuquerque, NM 87131, USA*

ABSTRACT

Motivation: Genome evolution is shaped not only by nucleotide substitutions but also by structural changes including gene and genome duplications, insertions and deletions and gene order rearrangements. Reconstruction of phylogeny based on gene order changes has been limited to cases where equal gene content or few deletions can be assumed. Since conserved duplicated regions are present in many genomes, the inference of duplication is needed in ancestral genome reconstruction.

Results: We apply GRAPPA-IR, a modified GRAPPA algorithm,to reconstruct ancestral chloroplast genomes containing duplicated genes. A test of GRAPPA-IR using six divergent chloroplast genomes from land plants and green algae recovers the phylogeny congruent with prior studies, while analyses that do not consider the duplications fail to obtain the accepted topology. The ancestral genome structure suggests that genome rearrangement in chloroplasts is probably limited by inverted repeats with a conserved core region. In addition, the boundaries of inverted repeats are hot spots for gene duplications or deletions.

Availability: The C source code for GRAPPA-IR is available upon request.

Contact: jtang@cse.sc.edu

INTRODUCTION

Mutations in a genome consist of not only base pair level changes but also events that alter the chromosome structure, such as inversions, duplications and deletions (Hurst *et al*., 2004; Kent *et al*., 2003). Ancestral gene sequence inference has led to significant predictions of protein functional shift and positive selection (Nei *et al*., 1997; Zhang and Nei, 1997; Muller *et al*., 2004). Comparisons of orthologous chromosomal segments showed heterogeneous rates of evolution of the X chromosome in human, mouse and rat (Gibbs *et al*., 2004). However, on the genome level the evolutionary change of genome structure is less well understood. Inference of ancestral genomes is limited to close related organisms with dense sample of sequences (Blanchette *et al*., 2004).

We take an alternative approach to study the genome structural changes by using the gene order data of fully sequenced chloroplast genomes. Chloroplasts are the green, photosynthetic organelles that originated from a free-living cyanobacteria-like ancestor (Raven and Allen, 2003). Chloroplast genomes have undergone significant downsizing while the genome structure has been maintained through over one billion years of endosymbiosis (Martin, 1998). Typical chloroplast genomes are circular single chromosomes consisting of 120-200 genes, which encode proteins, tRNAs, rRNAs and hypothetical open reading frames. Most chloroplast genomes consist of four distinct parts: two duplicated regions (inverted repeats, or *IR*) separated by a large single copy (*LSC*) and a small single copy (*SSC*) region. One common characteristic of the chloroplast IR is the presence of three rRNA genes (*rrn5s*, *rrn16s* and *rrn23s*, or *rrf*, *rrs*, and *rrl*), which are homologous to genes from the cyanobacteria *rrn* operon. The structure of chloroplast genomes of land plants is highly conserved, with almost collinear gene order, except for elevated level of rearrangements in specific lineages including green algae (Maul *et al*., 2002), conifers (Strauss *et al*., 1988) and members of the flowering plant families *Campanulaceae* (Cosner *et al*., 1997, 2004), *Geraniaceae* (Price *et al*., 1990) and *Fabaceae* (Perry *et al*., 2002). The gene content of chloroplast IRs vary greatly, largely due to the expansion and contraction of the IR at the IR-SC boundaries; this "ebb and flow" of the IR boundary has been observed even within a genus (Goulding *et al*., 1996; Plunkett and Downie, 2000). Chloroplast genomes of green algae (charophyte and chlorophyte algae) contain more variations of gene order and some are highly rearranged (Maul *et al*., 2002). Because of their compact size and the availability of conserved DNA probes, many chloroplast genomes have been mapped (Downie and Palmer, 1992) and 42 has been completely sequenced to date. Thus, chloroplast

*to whom correspondence should be addressed

genomes provide an ideal example for modeling genome rearrangements over a broad evolutionary time scale.

Gene order phylogeny was first proposed by David Sankoff (Sankoff *et al*., 1993); the software `BPAnalysis` implemented the algorithm using breakpoint distance, and applied it to animal mitochondrial genomes (Blanchette *et al*., 1999). It was not able to uniquely map the gene order changes and usually produced several tie trees, i.e. equally parsimonious trees regarding the optimization criterion. The inversion distance and inversion median were introduced to improve the phylogenetic accuracy and it has been implemented in the software `GRAPPA`. Extensive simulations showed that inversion medians were superior to breakpoint medians and the trees returned were more accurate using either distance-based or parsimony methods (Moret *et al*., 2001a; Bader *et al*., 2001; Moret *et al*., 2002a). Currently, `GRAPPA` (version 2.0) is able to estimate the phylogeny and true inversion medians using genomes with equal gene content (Moret *et al*., 2001b). A scaled-up version, `DCM-GRAPPA`, is able to estimate the gene-order phylogeny with apparently high accuracy for thousands of genomes, thus greatly increasing the power of genome phylogeny using large data sets (Tang and Moret, 2003).

Biologists are interested in simultaneous inference of ancestral genomes and the phylogeny from a set of known genomes. The inversion model is closer to the biological process of genome rearrangements, and inversion medians can be regarded as estimated ancestral gene orders. Here we attempt to reconstruct the ancestral chloroplast gene orders using selected completely sequenced genomes. Two challenges exist when we apply the algorithm for the ancestral genome reconstruction:

1. The algorithm needs to compute a phylogeny that includes heterogeneous branch lengths since the rate of chloroplast genome rearrangement varies significantly among lineages.
2. The algorithm also needs to handle gene duplications and deletions that lead to expansion and contraction of inverted repeats.

One version of `GRAPPA` is able to analyze genomes to infer phylogeny and ancestral gene orders using data sets with a limited number of deletions, but no duplication is allowed (Tang *et al*., 2004). To solve the problem, we have developed a new algorithm for chloroplast genomes that allows for the quadripartite structure (e.g., *LSC-IR-SSC-IR*) that is common to chloroplast genomes and other IR-containing DNAs. The assumption of the new approach is that inversions do not occur across inverted repeats, because the genome structure will be disrupted by such inversions that "flip" the orientation of the IRs. According to the model, a change of gene content within the IR region is mainly due to growth or shrinkage of the IR at the IR-SC boundaries. This approach is in agreement with observations of variable IR length and gene contents in most IR-containing chloroplasts.

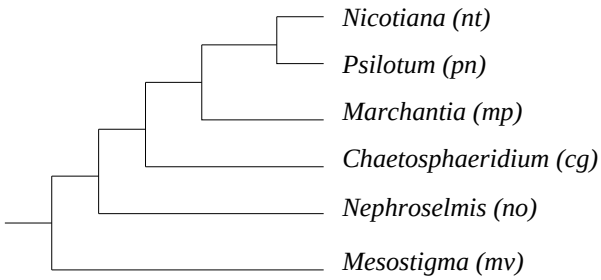

Fig. 1. The reference phylogeny of chloroplast genomes from land plants, green algae and a protist.

METHODS

The Dataset

Six chloroplast genomes representing major lineages of green plants and green algae were selected, all of which share the quadripartite structure. The organisms include *Nicotiana tobacum* (tobacco, *nt*), *Psilotum nudum* (whisk fern, *pn*), *Marchantia polymorpha* (liverwort, *mp*), *Chaetosphaeridium globosum* (a charophyte alga, *cp*), *Nephroselmis olivacea* (a chlorophyte alga, *no*) and *Mesostigma viride* (a photosynthetic protist, *mv*). A reference phylogenetic tree was constructed using the maximum parsimony and neighbor-joining methods on 50 concatenated proteins, and *Cyanophora paradoxa* proteins were used to root the tree (Figure 1). The reference tree is the same as the phylogeny by Lemieux *et al*. (2000), in which *Mesostimga* is basal to other green plants. An alterative phylogeny was published by Karol *et al*. (2001) based on maximum likelihood analysis of four chloroplast genes and including more algal taxa, in which *Mesostigma* is basal to charophytic green algae and sister to chlorophytic green algae.

We extracted 73 unique genes from the six genomes. Actual number of genes included in each genome ranges from 76 to 80 due to duplicated genes in the IR. The gene set includes 62 characterized protein-coding genes, 3 rRNAs, 7 tRNAs (identified by amino acid anticodons) and a hypothetical conserved open reading frame (*ycf1*). The encoding reflects the order and orientation of genes in the genome. Location of multi-exon genes was determined by the starting position of the first exon. In one case, the order of overlapping genes (*psbD-psbC*) was determined by the position of the start codon. The data set was then applied to a two-stage analysis to estimate ancestral gene orders.

Mapping Inverted Repeat to the Ancestral Genomes

We first consider the case when the gene content for each region of the genome (LSC, SSC and IR) are relatively conserved. When the genome is on a leaf (ie. it is an extant taxon and its gene order is known), we can easily determine the gene content for the LSC, SSC and IR regions through direct observation. However, since we do not know the gene order at each internal genome, we can only estimate the gene content for each region based on the assumption that all evolutionary events that alter the gene order are rare and that concurrent (i.e., parallel) changes in two children are less likely than a change in the parent. Thus, at each internal node, when the gene contents for a given region are known for for the two children, we face three possibilities of assigning a gene to the region:

1. If both children have gene g in the same region, then the parent had g in that region; otherwise, both children need to expand (or shrink) IRs and include g into that region, with a very low possibility.
2. If neither child has g, then g is absent in the parent. Since the genomes we test all share 73 unique genes, we do not consider this case in our study.
3. If g is located in different regions at the children, then it could be in either region of the parent. The two choices are equally likely if there is no further information from the phylogeny and we are left with an undetermined outcome for g.

If a gene is undetermined at some internal node, it can be resolved using an iterative improvement algorithm similar to the core algorithm in GRAPPA itself (the same method was used in Tang *et al*. (2004) for datasets with unequal gene content):

1. For each sibling pair of *leaves*, if a gene appears in the same region at both children, we place it in the same region at the parent (an internal node); If the gene appears in different region at the leaves, we mark its status as undetermined in the parent.
2. Starting from an arbitrary root, we carry out a depth-first search of the tree to propagate resolutions according to our standard rule – if two neighbors have the gene presented in the same region, the node will have it in that region too – and thus to resolve undetermined states through look-ahead and cost propagation.

Using the method above, we are able to determine the most likely gene contents for each possible tree (105 trees in this case). The estimated gene content for the internal nodes of the reference tree is presented in Figure 2.

Figure 2 shows that the gene contents of IR and SSC vary among the genomes. However, it also shows that the gene

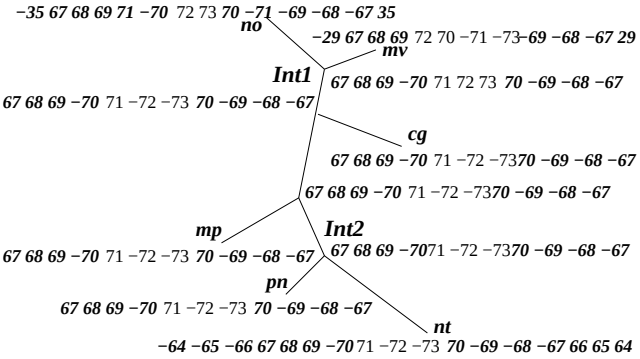

Fig. 2. Estimated gene contents for each region (only IR and SSC are shown).

content and gene order of IRs in the internal nodes are conserved, and four genes $(70, 71, 72, 73)$ are always kept together even if some of them are located in a different region. We assume that inversions did not cross the IR boundary in most chloroplast genomes, based on the observation that the gene contents in LSC and SSC are almost constant in the test data set.

Thus, we hypothesize the evolution of chloroplast genome structure as the following two steps:

1. The circular genome was divided into regions and inversions occurred in each region independently.
2. A segment from single copy regions was copied twice and joined to existing IRs, and the new genomes with longer IRs propagated. Alternatively, a segment was spliced out from one IR and new genomes with shorter IR propagated.

One should notice that the above two steps could happen several times along each edge. Based on this assumption, we could infer the possible evolutionary process from the internal node of *Int2* to *nt* and *Int1* to *no* , shown in Figure 3. This is a case of IR expansion.

For example, on the path that *Int2* was transformed to *nt*, each region of the genome underwent inversions independently, and three genes (64,65,66) moved close to the original IR boundary. After one duplication, the segment (-64 -65 -66) annealed with the original IR (67 68 69 -70) to form a new IR. If we remove the duplicates from the resulting IR, the gene contents of *Int*2 and *nt* would be identical.

From the observation above, we can further simplify the gene content of IR and SSC region, so that in the evolutionary path IR regions for all genomes (leaves and internal) contain gene (67 68 69), and the SSC regions contain gene (70 71 72 73) (although the gene order may be different). Furthermore, this operation treats all deletions or duplications as the last step in the path towards the observed gene orders and they do not need to be included in the scoring of internal genomes. Then it is possible to ignore the duplications and reduce

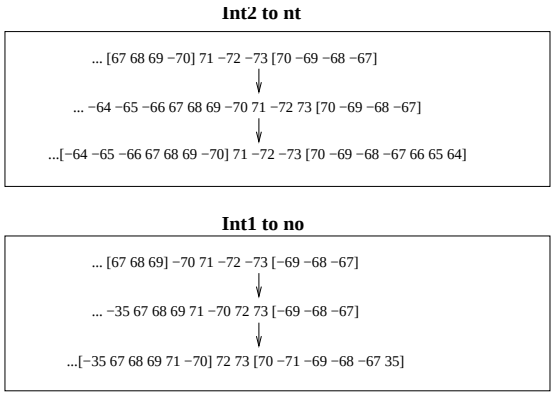

Fig. 3. Estimated evolutionary process from *Int2* to *nt* (top) and from *Int1* to *no* (bottom).

the problem to all leaf genomes of equal gene content. The simplified gene content map is shown in Figure 4:

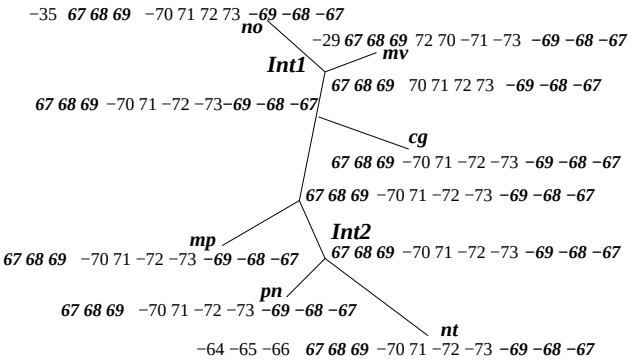

Fig. 4. Simplified gene content for IR and SSC. Each region contains the same genes before duplications.

Gene Order Phylogeny

We then reconstructed the phylogeny after the gene content of the ancestral genomes were determined. Since the gene contents were reduced to equal after the simplification step, it is feasible to use GRAPPA to infer an inversion phylogeny. If inversions are allowed to cross the boundaries of pre-determined IR and single copy regions, we can use the original GRAPPA to compute the ancestral gene orders and the phylogeny. However, this is unlikely, since we do not observe a single inversion involving both IR and genes in single copy regions. Thus, we have developed a new method, called GRAPPA-IR, that estimates inversions bounded by the boundary of IRs.

The new method still uses the exhaustive approach – it must score all possible trees to fin the one with the minimum number of inversions. To score a tree, it needs to solve the median problems of three genomes iteratively until no improvement can be found. However, this method differs from the original GRAPPA in the way it solves the median problems.

For three given genomes G_1, G_2 and G_3, solving the median problem is to find a genome G_0 that can minimize the sum of distances from itself to three given genomes. Since inversions do not cross IR boundaries, inversions in each region (LSC, SSC or IR) occur independently from other regions. In other words, the median problem can be divided into three sub-median problems, each of which is constructed from genes in the same region of the genome G_1, G_2 and G_3. The sub-median problems can be solved separately using available inversion median solvers (Caprara, 2001; Siepel and Moret, 2001).

Simulations

We set out to test the accuracy of GRAPPA-IR by simulations. For this purpose, we generated datasets of 6 and 10 genomes each and chose genomes of 78 genes (70 genes in the LSC region, 5 in SSC and 3 in IR), roughly in the range of our dataset described in the paper. We used a large range of evolutionary rates: let r denote the expected number of evolutionary events along an edge of the model tree, we used values of r in the range of $4-10$. The actual number of events along each edge is sampled from a uniform distribution on the set $\{1,2,\ldots,2r\}$. Given the model tree, we assigned the identity gene order to the root, and randomly generated gene orders for each node based on the edge length and the gene order of its parent assuming that inversions did not cross the IR boundaries. For each combination of parameter settings, we simulated 10 datasets and averaged the results.

Given an inferred tree (reconstructed phylogeny), we can assess the topological accuracy by *false positive* and *false negative* rates (Robinson and Foulds, 1981) with respect to the true tree. If an edge in the true tree is missing in the inferred tree, this edge is then called a *false negative* (FN). Similarly, a *false positive* edge (FP) appears in the inferred tree but not in the true tree.

We compared the new GRAPPA-IR to the original GRAPPA. We considered all trees with the minimum score given by both methods and took their strict consensus. Therefore, the trees returned by both methods need not to be fully resolved and they tend to have somewhat better rates for false positives (FP) than for false negatives (FN). Thus we report FN rates rather than FP rates or a single Robinson-Foulds score (Robinson and Foulds, 1981).

Figure 5 shows simulation results. This simulation indicates that GRAPPA-IR is clearly more accurate than the original GRAPPA for datasets with $r < 10$. Since our dataset falls into this category, we expect that using GRAPPA-IR will give us a better result.

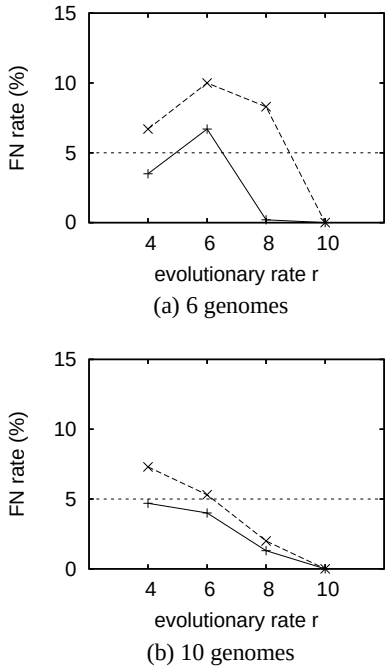

Fig. 5. False negative rate for GRAPPA-IR (solid line) and GRAPPA (dashed line) as a function of the evolutionary rate *r* for simulated datasets of 6 and 10 genomes. The horizontal line indicates the 5% error level, a typical threshold of acceptability for accuracy in phylogenetic reconstruction (Swofford *et al*., 1996).

RESULTS

We evaluated all trees for the six genomes using the new method. The best score returned is 76 after 100 min of computation on a PIV 3.4GHz workstation. The best tree (with score of 76) agrees with the reference tree (Figure 6). All the other trees are clearly worse, with scores no less than 78.

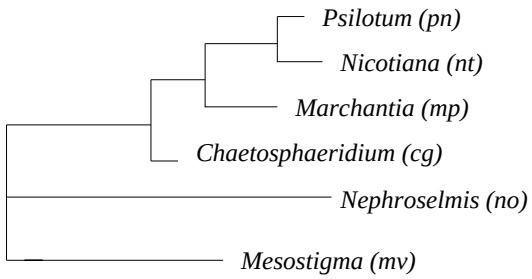

Fig. 6. The best tree obtained by GRAPPA-IR. The topology is the same as the reference tree.

We also tested the data set with the original GRAPPA, ignoring the region boundaries. The inference allowed inversions to occur across IR and single copy regions. The best tree obtained has the same score of 76, yet the topology (Figure 7) differs from the result of the previous test and is in conflict with the reference tree.

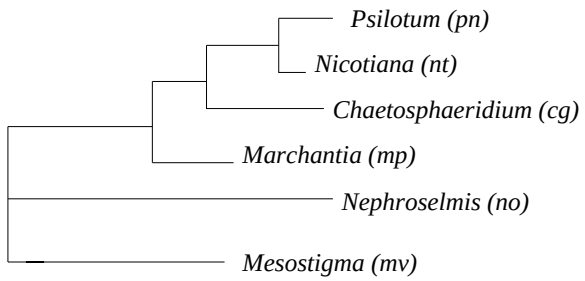

Fig. 7. The best tree obtained without the IR boundary limit. It misplaced two taxa, *mp* and *cg*, compared to the reference tree.

DISCUSSION

Ancestral Gene Cluster

We are able to reconstruct ancestral gene order from chloroplast genomes of land plants, green algae and a flagellate protist, spanning the history of green plant evolution. The ancestral chloroplast genome of land plants and algae contains IR, which is consistent with the hypothesis that IR is a feature derived early in chloroplast endosymbiosis (Palmer, 1985). Although the sequenced cyanobacteria *Nostoc* and *Synechococcus* do not maintain *rrn*-containing IRs, if other cyanobacteria are identified with structures similar to the chloroplast IR, then it would suggest an even earlier origin for this structure. In addition, the ancestral IR contains the same gene content to that of *Mesostigma*, which agrees with the observation that *Mesostigma* chloroplast genome encodes several ancestral gene clusters (Lemieux *et al*., 2000). By comparison of ancestral gene orders to the extant genomes, it is possible to test formally the evolutionary force of gene order changes. For example, ancestral gene clusters may be more likely to be maintained if they share related function and are under constraints (Stoebe and Kowallik, 1999), while continuous rearrangements would lead to break down of less constrained clusters.

IR and Genome Stability

The gene content of IR varies across land plants, even in a single genus or family (Goulding *et al*., 1996). IR motifs are thought to mediate intragenic inversions by forming a stem-loop structure (Palmer and Thompson, 1982; Graham and Olmstead, 2000), and homologous recombinations between the two repeats are frequent (Palmer, 1985). In a single chloroplast, hundreds of copies of chloroplast DNA co-exist as circular monomer, dimer and linear chromosomes (Bendich and Smith, 1990). In the cellular endosymbiosis environment, the selection on accuracy of replication may have been

relaxed to the degree that unequal recombination and replication slippage contribute to the expansion or shrinkage of IRs. Short repeat motifs may facilitate inter-molecular recombination and create diversity of chloroplast genomes in a population (Kawata *et al*., 1997). On the other hand, the intra-molecular recombination homogenize the sequence of IR and thus the particular IR size and the gene content are maintained. The two counteracting phenomena may have played important roles in shaping the current diversity of chloroplast genome gene orders.

We find that wrong gene order phylogeny are recovered without the IR boundary information. This suggests that maintenance of IR is necessary in the evolution of chloroplast genomes in most of the cases. It supports the hypothesis that IR provides an insulation mechanism that stabilizes the genome structure, and the genes in single copy regions do not commute across the IR (Palmer, 1985). This agrees with the observation that gene rearrangements are more frequent in chloroplast genomes without IR (Palmer and Thompson, 1982). However, some genomes with residual IRs but infrequent gene movements between single copy regions compared to related lineages do not conform to the hypothesis (Cosner *et al*., 2004). Future experimental studies on highly rearranged chloroplast genomes in the green algae may shed light on the maintenance of IR and genome rearrangements.

Comparison to Other Methods

A similar approach used for human and mouse genome comparison, `GRIMM`, showed the optimal sorting of X chromosomes by at least 7 inversions (Pevzner and Tesler, 2003). This is moderate amount of changes compared to the level we observe in many chloroplast genomes. If duplications and deletions are considered in a finer scale, the process will be much complex, as suggested by the reconstruction of one 1.1 Mb region in the eutherian mammal ancestor (Blanchette *et al*., 2004). Extensive tests show that trees returned by `GRAPPA` are superior to those returned by other methods (Moret *et al*., 2002a). The closely related package of Pevzner's group, `MGR` (Bourque and Pevzner, 2002), is the only one that approaches its accuracy. It uses `GRIMM` and enables analysis of multi-chromosomal genomes. In the single chromosomal case (such as organelle genomes), our software consistently achieves high accuracy and efficiency.

CONCLUSIONS

We have implemented a new method, `GRAPPA-IR`, to infer ancestral gene orders with duplications. Tests on a real data set and simulations show accurate recovery of the genome phylogeny, as well as fast inference of ancestral gene orders. This provides new insight into the genome evolutionary process. There still lie challenges to apply the method to large datasets of bacterial or eukaryotic genomes. A combination of disk-covering method and other approaches may scale up the capability to infer ancestral gene order for large genomes.

ACKNOWLEDGEMENTS

The work is supported by NSF grants DBI 01-15684, DEB 01-20709 to C.D., grants ANI 02-03584, EF 03-31654, IIS 01-13095, IIS 01-21377, DEB 01-20709 to B.M. and by the Dept. of Computer Science and Engineering at U. of South Carolina to J.T..

REFERENCES

Bader,D.A., Moret,B.M.E. and Yan,M. (2001) A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J. Comput. Biol.*, **8**, 483–491.

Bendich,A.J. and Smith,S.B. (1990) Structure of chloroplast and mitochondrial DNAs. *Curr. Genet.*, **17**, 421–425.

Blanchette,M., Kunisawa,T. and Sankoff,D. (1999) Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.*, **49**, 193–203.

Blanchette,M., Green,E.D., Miller,W. and Haussler,D. (2004) Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res.*, **14**, 2412–2423.

Bourque,G. and Pevzner,P.A. (2002) Genome-Scale Evolution: Reconstructing Gene Orders in the Ancestral Species. *Genome Res.*, **12**, 26–36.

Caprara,A. (2001) On the practical solution of the reversal median problem. In *Proc. 1st Int'l Workshop Algs. in Bioinformatics (WABI'01), Lecture Notes in Computer Science*, **2149**, 238–251.

Cosner,M.E., Jansen,R.K., Palmer,J.D. and Downie,S.R.(1997) The highly rearranged chloroplast genome of Trachelium caeruleum (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. *Curr. Genet.*, **31**, 419–429.

Cosner,M.E., Raubeson,L.A. and Jansen,R.K. (2004) Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. *BMC. Evol. Biol.*, **4**, 27.

Downie,S.R. and Palmer,J.D. (1992) Restriction Site Mapping of the Chloroplast DNA Inverted Repeat - a Molecular Phylogeny of the Asteridae. *Annals of the Missouri Botanical Garden*, **79**, 266–283.

Gibbs,R.A., Weinstock,G.M., Metzker,M.L., Muzny,D.M., Sodergren,E.J., Scherer,S., Scott,G., Steffen,D., Worley,K.C., Burch,P.E. et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.

Goulding,S.E., Olmstead,R.G., Morden,C.W. and Wolfe,K.H. (1996) Ebb and fbw of the chloroplast inverted repeat. *Mol. Gen. Genet.*, **252**, 195–206.

Graham,S.W. and Olmstead,R.G. (2000) Evolutionary signifi cance of an unusual chloroplast DNA inversion found in two basal angiosperm lineages. *Curr. Genet.*, **37**, 183–188.

Hannenhalli,S. and Pevzner,P.A. (1999) Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. *J. ACM*, **46**, 1–27.

Hurst,L.D., Pal,C. and Lercher,M.J. (2004) The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.*, **5**, 299–310.

Karol,K.G., McCourt,R.M., Cimino,M.T. and Delwiche,C.F. (2001) The closest living relatives of land plants. *Science*, **294**, 2351–2353.

Kawata,M., Harada,T., Shimamoto,Y., Oono,K. and Takaiwa,F. (1997) Short inverted repeats function as hotspots of intermolecular recombination giving rise to oligomers of deleted plastid DNAs (ptDNAs). *Curr. Genet.*, **31**, 179–184.

Kent,W.J., Baertsch,R., Hinrichs,A., Miller,W. and Haussler,D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA*, **100**, 11484–11489.

Lemieux,C., Otis,C. and Turmel M. (2000) Ancestral chloroplast genome in Mesostigma viride reveals an early branch of green plant evolution. *Nature*, **403**, 649–652.

Martin,W., Stoebe,B., Goremykin,V., Hapsmann,S., Hasegawa,M. and Kowallik,K.V. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature*, **393**, 162–165.

Maul,J.E., Lilly,J.W., Cui,L., dePamphilis,C.W., Miller,W., Harris,E.H. and Stern,D.B. (2002) The Chlamydomonas reinhardtii plastid chromosome: islands of genes in a sea of repeats. *Plant Cell*, **14**, 2659–2679.

Moret,B.M.E., Wyman,S., Bader,D.A., Warnow,T. and Yan,M. (2001a) A new implementation and detailed study of breakpoint analysis. In *6th Pac. Symp. Biocomput. (PSB2001)*. Big Island, Hawaii, January 2001, pp.583–594.

Moret,B.M.E., Wang,L.S., Warnow,T. and Wyman,S.K. (2001b) New approaches for reconstructing phylogenies from gene order data. *Bioinformatics*, **17** Suppl 1, S165–173.

Moret,B.M.E., Tang,J., Wang,L-S. and Warnow,T. (2002a) Steps toward accurate reconstructions of phylogenies from gene-order data *J. Comp. Syst. Sci*, **65**, 508–525

Moret,B.M.E., Siepel,A.C., Tang,J. and Liu,T. (2002b) Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. *Lecture Notes in Computer Science*, **2452**, 521–536.

Muller,K., Borsch,T., Legendre,L., Porembski,S., Theisen,I. and Barthlott,W. (2004) Evolution of carnivory in Lentibulariaceae and the Lamiales. *Plant Biol. (Stuttg)*, **6**, 477–490.

Nei,M., Zhang,J. and Yokoyama,S. (1997) Color vision of ancestral organisms of higher primates. *Mol. Biol. Evol.*, **14**, 611–618.

Palmer,J.D. and Thompson,W.F. (1982) Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell*, **29**, 537–550.

Palmer,J.D. (1985) Evolution of Chloroplast and Mitochondrial DNA in Plants and Algae. In *Molecular Evolutionary Genetics*. Edited by MacIntyre, R.J.. New York: Plenum Press; pp. 131–240.

Perry,A.S., Brennan,S., Murphy,D.J., Kavanagh,T.A. and Wolfe,K.H. (2002) Evolutionary re-organisation of a large operon in adzuki bean chloroplast DNA caused by inverted repeat movement. *DNA Res.*, **9**, 157–162.

Pevzner,P. and Tesler,G. (2003) Genome Rearrangements in Mammalian Evolution: Lessons From Human and Mouse Genomes. *Genome Res.*, **13**, 37–45.

Plunkett,G.M. and Downie,S.R. (2000) Expansion and Contraction of the Chloroplast Inverted Repeat in Apiaceae Subfamily Apioideae. *Syst. Bot.*, **25**, 648-667.

Price,R.A., Calie,P.J., Downie,S.R., Logsdon,J.M. and Palmer,J.D. (1990) Chloroplast DNA variation in the Geraniaceae - a preliminary report. In *Proc. int Geraniaceae symp.* Edited by Vorster P. University of Stellenbosch; Monvilla, South Africa, 1990. pp. 235-244.

Raven,J.A. and Allen,J.F. (2003) Genomics and chloroplast evolution: what did cyanobacteria do for plants? *Genome Biol.*, **4**, 209.

Robinson,D.R. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**, 131–147.

Sankoff,D. (1993) Analytical approaches to genomic evolution. *Biochimie*, **75**, 409–413.

Siepel,A.C. and Moret,B.M.E. (2001) Finding an optimal inversion median: experimental results. In *Proc. 1st Int'l Workshop Algs. in Bioinformatics (WABI'01), Lecture Notes in Computer Science*, **2149**, 189–203.

Stoebe,B. and Kowallik,K.V. (1999) Gene-cluster analysis of chloroplast genomics *Trends Genet.*, **15**, 344–347.

Strauss,S.H., Palmer,J.D., Howe,G.T., Doerksen,A.H., Edwards,H., Jorgensen,R.A. and Thompson, W. F. (1988) Chloroplast genomes of two conifers lack a large inverted repeat and are extensively rearranged. *Proc. Natl. Acad. Sci. USA*, **85**, 3898–3902.

Swofford,D.L., Olson,G., Waddell,P. and Hillis,D.M. (1996) Phylogenetic inference. In Hillis,D.M., Moritz,M. and Mable, B. (eds), *Molecular Systematics, 2nd ed.*, Sinauer Associates, Sunderland, pp. 407 –514.

Tang,J. and Moret,B.M.E. (2003) Scaling up accurate phylogenetic reconstruction from gene-order data. *Bioinformatics*, **19** Suppl 1, i305–312.

Tang,J., Moret,B.M.E., Cui,L. and dePamphilis,C.W. (2004) Phylogenetic Reconstruction from Arabitrary Gene-Order Data. In *Proc. 4th IEEE Symp. on Bioinfor. and Bioeng. (BIBE'04)*. Taichung, Taiwan. May 2004, pp. 592–599.

Zhang,J. and Nei,M. (1997) Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol.*, **44** Suppl 1, S139–146.